From: Chien Tai

Approach:

I started out by importing all the necessary python statistical packages to be used in the analysis including numpy, pandas, cross validation, logistic regression and plotting..etc.

The training data was downloaded to a local drive and read in as a dataframe using the read_csv function under pandas.

#1. The next step is to start the data exploration process and impute missing values as needed. Using the info() function, we observed that Age only has 714 non-null entries. I used scatter matrix via pandas to see if there is any correlation between Age and other variables and decide how to impute the Age missing values. From the scatter matrix, Age appears to be correlated with Parch and SibSp. The correlation appears to be strongest with SibSp. This makes intuitive sense in that the more sibling/spouses the individual is traveling with, the more likely the person would be of younger age as most younger individuals are traveling with their family. Subsequently, I decided to calculate the median age by number of Siblings and Spouse the person has and impute the missing value that way. (Another possibility is to impute it using median age by gender and/or Parch)

#2 a. In this step, we will try to clean the data and use logistic regression to predict whether a passenger would survive. We started out by looking at the scatter plot to see whether survival rate is linked to a particular variable. Panda Pivot tables are also used to look at survival rate by variables. We found that Survival rates are correlated with Pclass, Age, SibSp, Parch, Fare and Gender. Survival rate by gender is pasted below to illustrate the survival rate difference.

	mean
Sex	**Survived**
female	.7420380
male	0.188908

We also recode Sex to transform it into a Boolean 0 and 1 variable called gender. Irrelevant columns were dropped before we fit a logistic regression. Random seed is specified to make sure results are re-producible.

2b. A few things were tried: normalizing the data, testing out optimal regularization parameter. Normalizing the data does not appear to change the model fit.

When the following regularization parameters were tested, we observed that C = 1 gives us the best cross validation score. (the array below shows, C value, cross validation score]

```
[(1000, 0.78787878787878773), (100, 0.78787878787878773), (10, 0.7890
0112233445563), (1, 0.79349046015712688), (0.1, 0.77665544332211001),
  (0.01, 0.70594837261503918), (0.001, 0.68686868686868685), (0.0001,
 0.66666666666666663)]
```

The model coefficient output by variable is as followed:

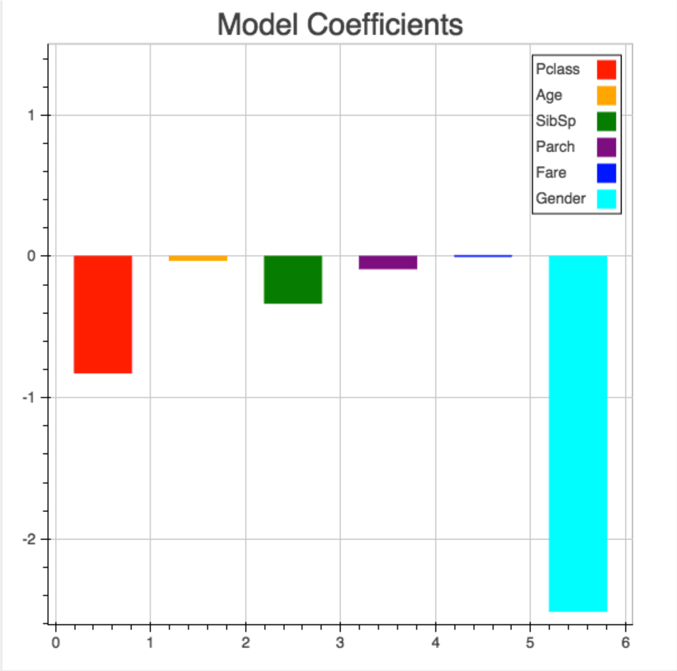
```
Index([u'Pclass', u'Age', u'SibSp', u'Parch', u'Fare', u'Gender'], dtyp
e='object')
```

```
[-0.6560228  -0.01691447 -0.38819441 -0.13876083  0.00756292 -2.5639129
9]
```

Confusion matrix is used below to evaluate model effectiveness.

```
Results of Logistic Regression:
                        Actual Class 0  Actual Class 1
 Predicted Class 0             115              13
 Predicted Class 1              33              62
Precision: 0.826666666667
Recall: 0.652631578947
```

2c. The coefficient by variable is graphed below. We can tell that Gender, Pclass and SibSp appears to have the highest coefficient. Intuitively speaking, we know at the time of the ship wreck (early 1900's), women, children, and 'upper class' individuals were given the priority to board the lifeboat first. At the same time, when one is traveling with siblings and spouses, they might also be more likely to be traveling with children and hence has higher survival predictive power.



2d. The python code should be flexible enough to work with a different dataset.

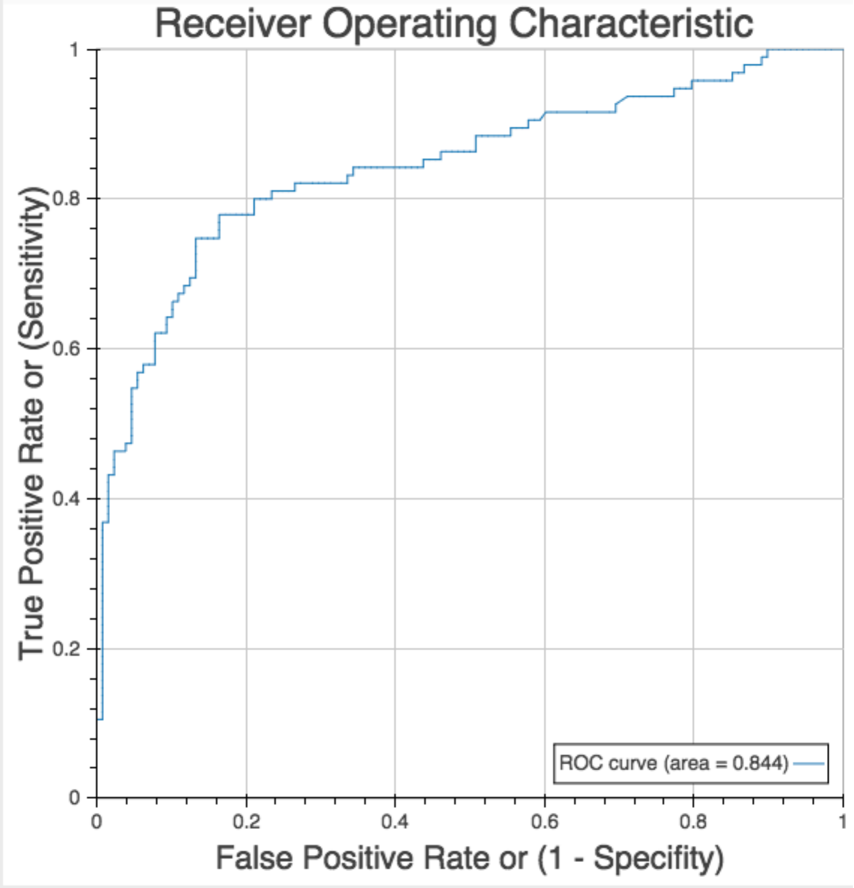
3. Cross validation technique was used to look at model prediction. We tested out the following cross validation folds to see whether it significant impact cross validation outcome. (k_list = [3, 5, 10, 20, 50, 100]) The result is shown below.

```
[(3, 0.79349046015712688), (5, 0.78789553962763859), (10, 0.793503291
34037001), (20, 0.79343983311374633), (50, 0.79480392156862734), (10
0, 0.79366666666666674)]
```

While 50 folds appears to have the highest cross-validation score. All the other ones produce similar results too. In this case 10 is chosen as it has a fairly high cross validation score and has been an industry standard. Generally speaking, the logistic regression has a cross-validation score close to 80%. The model is performing relatively well.

4a & b

The ROC curve is produced and pasted below. The area under curve (AUC) in this model is 0.844.



4c. The ROC has a shape that's desirable in that it curves towards the upper left quadrant relatively to the 45 degree line. The AUC is also favorable. This result is achieved since I had selected a few variables with good predicted power on survival rates, the model has also been evaluated to choose optimal regularization parameter. Previously, we've seen the model produced relatively high sensitivity given low false positive rates. This is more predictive than simply guessing the outcome randomly.

4d. A potential next step is to change the default threshold in logistic regression classification to get to a point where we can be closest to the upper left point on the ROC.

4e. The default threshold for logistic regression is 0.5. That currently correspond to a ROC point at ~ (0.1, 0.65). Looking at the ROC curve above, we observe that the optimal point should be close to (0.2, 0.8). The optimal point would correspond to a threshold close to 0.33. It would make sense to re-purpose the logistic model with a 0.33 threshold and see how it performs on cross-validation and confusion matrix.